Root Cause Analysis for PBI000000000045:

D0ora2 hardware incident response delays on 21 August 2009

Report submitted 14 September 2009

(CD-doc-3392)

Summary of Incident

The offline database system for the D0 experiment, d0ora2, experienced a service disruption in the early hours of Friday, August 21, 2009. First report of trouble (that has come to light so far) was an email to the d0db-support mail list around 1:55 AM. This list is not monitored 24x7. An Incident ticket was also generated around this time through the Requester Console, but that is not the proper procedure for reporting a service disruption to a system with 24x7 support. The DBAs did respond to the email when they came in that morning, determined they could not access the system and contacted the system administrator. The administrators also were unable to access the system and rebooting was unsuccessful. A hardware support request process was started with Decision One/Sun. Initially a misdiagnosis by the Sun engineer resulted in wrong part being ordered (which had a long delivery time). Upon delivery of the part, a Decision One technician determine it was the wrong part based on the diagnostics and indicated a different part was needed. Decision One failed to call back system administrator in the 2 hour response window, but did respond later when contacted directly by local system administrator. The administrator was told they had the part (around 7:30 PM) but later learned from an engineer at 9:00 PM the part was not ordered, he would do it ten and it would take 90 more minutes. In the end the service was down more than 24 hours, and several of these delays seem unnecessary or unnecessarily long.

List of related Incident tickets:

- INC000000008527
- INC000000008542

Background Concepts

The following points are useful in understanding the nature and impact of this problem:

- D0ora2 is the production offline database server for the D0 experiment. It is a Sun server running Solaris.
- D0ora2 is a 24x7 supported system critical for offline physics analysis and transfer of online raw data files for the D0 experiment.
- Hardware Maintenance contract is with Decision One. Sun contracts out hardware service to Decision One.

- Normally the D0 experiment control room is in 24x7 operations, but due to a planned and lengthy accelerator shutdown staffing for things like monitoring was reduced, and hardware maintenance activities tend to be clustered during these periods.
- New Incident Management tool and procedures were put into production earlier this year. There is not yet uniform familiarity and usage of the tool throughout the Incident investigation process among all support groups.
- Support personnel were in the process of commissioning a replacement system for D0ora2, called D0ora3, and proper completion of commissioning this new system required D0ora2 to be operational.
- Maintenance of the contract with Decision One is in the process of changing ownership within the Computing Division and new ownership is not fully aware of the details in the contract.
- Currently the hardware request form does not create a ticket in the Incident Management system or in the separate hardware ticket system used by Decision One. Entry into the hardware ticket system is a manual process. An investigation into automating this process is planned.
- Sun Case #71515806

Timeline

This is a partial timeline for the service disruption and restoration.

Friday, August 21st

01:55 Qizhong Li reports service disruption via email to d0db-support email list and through the Requester Console.

08:23 Steve White takes a look and reports a DBA needs to investigate.

08:40 Jared Platson reports d0ora2 has connection problems and Maurine cannot get to console.

09:17 Maurine Mihalek initiates a reboot of the system.

09:38 Maurine Mihalek reports system is still trying to reboot and she has called Sun to open a ticket.

09:44 Maurine Mihalek calls Service Desk and opens a hardware request through the web interface. Also leaves voice messages with Service Desk staff and the Incident Manager.

10:11 Maurine Mihalek reports Sun claims there is a problem with an internal Fibre Channel HBA controlling all the internal disks.

10:19 Service Desk calls Maurine to say they will process hardware request.

10:42 Mike Militec from Decision One calls to say best guess for part is 2 hours, he still does not see ticket numbers. Maurine provides numbers.

11:51 Tom Ackenhusen sends a status update.

12:40 Maurine Mihalek calls Mike Militec from Decision One about part status but has to leave a message.

14:20 Maurine Mihalek relays message that part is being driven from Racine, Wisconsin with an estimated time of arrival around 17:00.

16:49 Tom Ackenhusen sends another status update.

18:07 Maurine Mihalek reports the wrong part was ordered, and the initial diagnosis from Sun was wrong. New diagnosis from Sun claims the motherboard needs to be replaced.

19:29 Maurine Mihalek reports having to contact Sun for status and learned no engineer had been assigned yet. She also reports that a second look at switching over to d0ora3 concludes that is not feasible at this time.

21:01 Maurine Mihalek reports part has not yet been ordered but engineer is going that now, estimating a 90-minute delivery.

Saturday, August 22nd

01:23 Maurine reports that the server is back up and will contact DBAs for Oracle service restoration.

02:01 Svetlana Lebedeva reports that the database servers are up and running.

02:16 Maurine Mihalek reports the system is panicking and going down.

02:41 Svetlana Lebedeva reports system is back up and databases started automatically. Anil Garg replies a few minutes later that he manually started the servers after the system came up.

04:25 Maurine Mihalek reports that Sun believes the panic was likely caused by components being disturbed when the motherboard was replaced. Memory errors are being corrected but could lead to another panic if too many seen to quickly.

Analysis

The hardware service disruption lasted for roughly 24 hours spanning from the early morning hours of Friday, August 21st to Saturday, August 22nd. A system disk hardware failure initially resulted in loss of the ability to open TCP/IP connections but allowed network pings to succeed. The disruption was first noticed on the customer side, due to failures in database query requests, and reported via email to a database support list and through the Requester Console. Although the service is listed as being under 24x7 support, neither of these reporting methods is appropriate for initiating 24x7 support. An automated query normally is monitored

in the experiment control room, but was disabled at the time. However even if enabled and watched, it would still have required a manual intervention by the control room staff to request a page.

On the service provider side, none of the automated monitoring checks were designed to trigger on this failure mode.

In the absence of an automated trigger, the initial service disruption report via normal business hours channels resulted in an unnecessary delay of roughly 6 hours, or about one quarter of the overall physical service disruption period. The dominant fraction of the time was due to performance with respect to underlying contracts for hardware and software support. Although Sun contracts out hardware support through Decision One, they initially attempted to diagnose the disruption remotely and without contacting Decision One. This decision resulted in a misdiagnosis of the failure and the ordering of the wrong replacement part (one that apparently does not even fit in the machine). Based on the emails and notes of our support staff, there are strong indications that communications breakdowns at Decision One and Sun also were a factor. The performance of these two organizations seems inconsistent with reasonable expectations, although details of the actual contracts were not available for the root cause analysis.

Internally and customer facing there were numerous updates provided throughout the service disruption period once support staff were aware of the situation. However usage of the Incident Management system was minimal at best, and neither of the two Incident tickets were updated while the disruption was in progress. It is noted that both internal support organizations involved in this Incident investigation reside in the Lab and Science Core Services quadrant, and this quadrant is spearheading the ITIL effort within the division and laboratory.

Based on the impact and urgency of this disruption, it could have been declared a Critical Incident and handled through the Critical Incident Management process. The Critical Incident Manager would have coordinated internal and external communications. It is however noted that the role of Critical Incident Manager has not yet been assigned and this would have diminished the value gained from the process.

The Fishbone diagram shown in Figure 1 identifies two areas where improved monitoring could trigger on future Incidents of a similar nature. A simple system command executed remotely via ssh would provide an enhanced test of system operations compared to a network ping. Additionally, from a service perspective, an automated query integrated with the paging system would be a good check on basic functioning of the service.

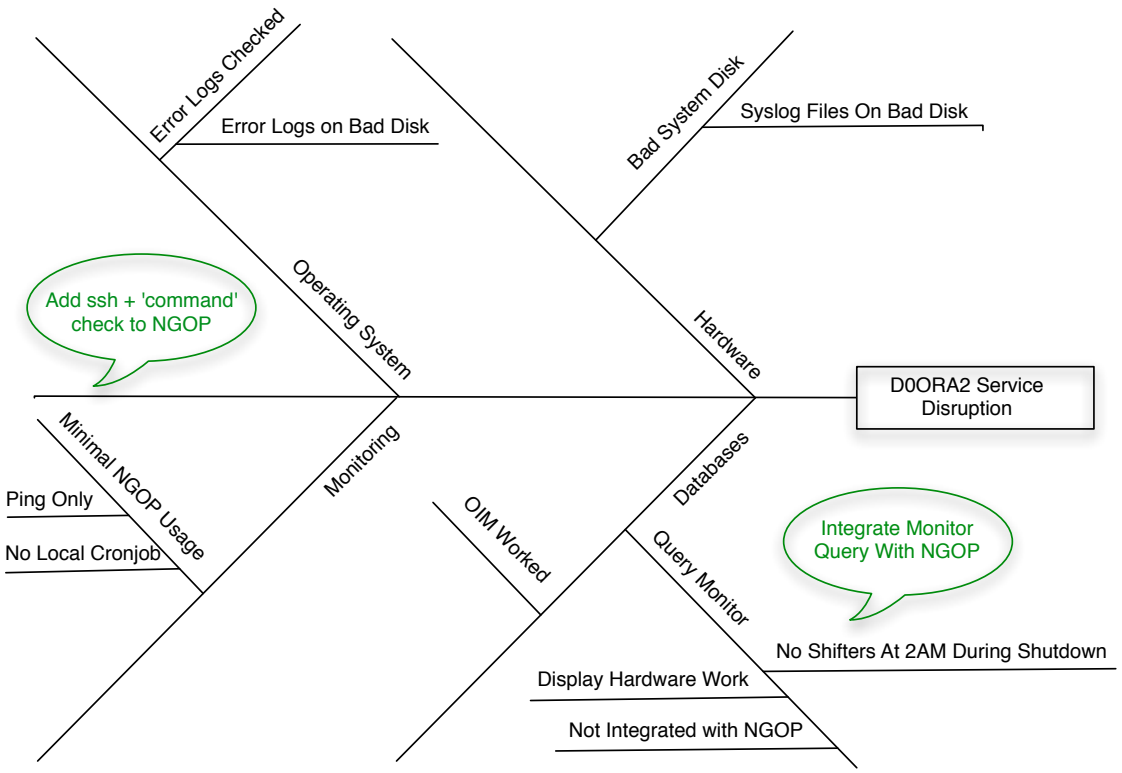

Figure 1: Fishbone diagram for D0ora2 hardware service disruption root cause analysis. Green dialog bubbles indicate ways to address weaknesses identified.

Direct Cause

System disk hardware failure resulted in unresponsive behavior from some, but not all, system commands and resulted in database query failures from external systems.

Contributing Factors

A number of factors contributed to the failure to alert support personnel in a timely manner of the initial service disruption or to the timely restoration of the service once support personnel were notified. There is no significance to the ordering of these contributing factors.

- There exists a query of the database that is normally monitored on a display in the control room. The query was disabled to work on the monitor display system and not immediately re-enabled once the system was back in service.
- Customer who first observed a service disruption did not follow proper procedures for reporting a disruption of a 24x7 service during off hours.
- Accelerator shutdown resulted in minimal staffing during off hours and thus would not have seen a query failure had that process been enabled.
- Automated query of the offline database was disabled while maintenance was performed on the display host. However re-enabling of the query was not done once the system was returned to service.

- Initial misdiagnosis of the hardware failure lead to a delay due to part availability and the subsequent further investigation into the outage.
- Further misdiagnosis of the underlying cause and miscommunications in ordering of additional parts resulted in further delays.
- Decision One did not physically inspect the machine and attempt to confirm diagnosis, even though part delivery was estimated to take many hours, until after the first part arrived.
- Sun did not involve Decision One early in the process even though their initial diagnosis indicated a hardware failure.

Root Cause

Insufficient automated monitoring of query functionality from external systems failed to catch service disruption and notify support personnel.

Observations & Comments

A number of observations and comments were collected during the initial Problem logging and investigation, or in the subsequent root cause analysis meeting. There is no significance to the ordering of these observations and comments.

1. The information provided by the Decision One management with respect to the timeline of notifications on this service disruption are not in agreement with the detailed log entries provided by the system administrator.
2. Support personnel are not seeing clear management guidance on following the Incident Management processes and tool usage.
3. Training on usage of the updated Incident Management tool may be insufficient or not adequately communicated.
4. Coordination and management of Decision One and Sun support had to be done by system administrator, but should have been handled by the Service Desk or in the case of a Critical Incident, the Critical Incident Manager.
5. Some DBAs reported not seeing email on INC000000008542 being assigned to their group. However investigation in both Remedy audit logs and in mail server and IMAP logs indicate email was sent by Remedy and received by the IMAP server for all DBAs on IMAP servers. One person on the DBA list has email forwarded elsewhere but he reported seeing the email.

Recommendations

Based on the Problem investigation and root cause analysis for the D0ora2 hardware service disruption, the following list of recommendations are made for preventing future delays in timely notification and restoration of service disruptions to this system.

1. The production offline database server, formerly D0ora2 and now D0ora3, has 24x7 support. Therefore any service disruption should be reported according to the policies and procedures agreed upon between the customer and service provider for services with 24x7 support. A failure to follow the

policies and procedures could result in an unnecessarily long service disruption beyond the control of the service provider.

2. There are serious concerns with the handling of the hardware service disruption by Decision One and Sun. A formal review of the contract should be conducted within the Computing Division to understand service level expectations, procedures and metrics for assessing performance of the service provided by the hardware support contractor. Preferably this review would include both management and appropriate support personnel.
3. Lab And Scientific and Core Services management at all levels should clearly communicate Incident Management policy, procedures and expectations, including proper usage of the Incident Management tool, to all members of the quadrant. Department heads and group leaders should be accountable for members of their organization following these policies and procedures.
4. An automated query of the offline database should be integrated with NGOP and configured to page support personnel through the Incident Management system.
5. A check should be added to NGOP monitoring that invokes a command on the database server using ssh. This command should be simple but constructed at least to result in some access to the system disk.
6. The ITIL sponsors should assign the Critical Incident Manager role and implement this process within the organization as soon as possible.

Root Cause Analysis Committee

The following people served on the RCA committee (4 September 2009):

Gerald Guglielmo (lead)

Svetlana Lebedeva

Adam Lyon

Maurine Mihalek

Jared Platson

Carin Sinclair

Nelly Stanfield